When Is a Phylogenetic Test Good Enough?

John W. *Wenzel*

Department of Entomology, Ohio State University, Columbus, OH., 43210, USA

ABSTRACT

Cladistic viewpoints have not been widely appreciated during the rapid growth of interest in phylogenetic tests of evolutionary scenarios. General opinion sometimes contrasts with cladistic perspectives with respect to the nature or severity of certain problems. Cladistic views are straightforward, if somewhat counter-intuitive in certain constituencies. The tree that best summarizes the data is always the most parsimonious tree (or consensus of most parsimonious trees) and nodes on this tree should not be dissolved based on low numerical support, such as bootstrap values. Resolution of the consensus tree can be a problem when data are inadequate. Successive approximations weighting to derive better resolution is consistent with the cladistic paradigm in that congruence among characters determines the relative weights of the characters. This process is recursive, but not circular. In contrast, techniques such as maximum likelihood should not be used to derive a more resolved tree because that tree will not be based upon the data alone; results are biased according to (sometimes extensive) *a priori* dictations of the probable path of evolution and they will not reveal patterns incongruent with the initial assumptions. Construction of maximum likelihood trees relies on process theories unrelated to, and perhaps uninformative for, adaptive traits of interest. The characters of interest should be included in the final analysis because they are the data most relevant to the analysis. Claims that eliminating them improves independence or that including them leads to circular reasoning are incorrect both logically and empirically. Eliminating characters because they are expected to show high homoplasy is an unacceptable *ad hoc* protection of an hypothesis from a legitimate test. Most statistical treatments require unnecessary and often unsupportable assumptions regarding the process of evolution and expected distribution of traits. Character data and their appearance on the most parsimonious tree are preferred because they are the most free of assumptions and remain the most closely tied to the actual data.

RÉSUMÉ

Qu'est ce qu'un bon test phylogénétique?

Le point de vue de la cladistique n'a pas été suffisamment pris en compte à l'occasion de l'intérêt croissant manifesté pour les tests phylogénétiques de scénarios évolutifs. L'opinion générale est souvent en désaccord avec les perspectives cladistes au sujet de la nature ou de l'importance de certains problèmes. Les points de vue cladistes sont clairs, même si ils paraissent aller à l'encontre de l'intuition dans certains domaines. L'arbre qui rend compte le mieux des données est toujours l'arbre le plus parcimonieux (ou le consensus des arbres les plus parcimonieux) et des noeuds de cet arbre ne doivent pas être abandonnés parce qu'ils présentent de faibles valeurs de paramètres numériques, telles que des valeurs de bootstrap. La résolution d'un arbre consensus peut constituer un problème quand les données sont inadéquates. La pondération successive appliquée à l'amélioration de la résolution est en accord avec le paradigme cladistique : c'est bien la congruence entre les caractères qui détermine les poids relatifs de ces caractères. La procédure est récursive et non pas circulaire. A l'opposé, les techniques telles que le maximum de vraisemblance ne devraient pas être utilisées pour obtenir un arbre mieux résolu parce que cet arbre n'est plus basé sur les seules données ; les résultats incorporent alors de manière parfois importante des opinions *a priori* sur le

Wenzel, J. W., 1997. — When is a phylogenetic test good enough? *In*: Grandcolas, P. (ed.), The Origin of Biodiversity in Insects: Phylogenetic Tests of Evolutionary Scenarios. *Mém. Mus. natn. Hist. nat.*, **173** : 31-45. Paris ISBN : 2-85653-508-9.

voies évolutives les plus probables et ils ne révéleront plus des schémas incongruents avec les hypothèses de départ. La construction d'arbres de maximum de vraisemblance repose sur des processus théoriques sans rapport, voire non informatifs, avec les traits adaptatifs à l'étude. Les caractères à l'étude doivent être inclus dans l'analyse finale parce que ce sont les données les plus pertinentes en regard de cette analyse. Certains ont prétendu que leur élimination améliorerait l'indépendance ou que leur inclusion conduirait au raisonnement circulaire : ces assertions sont incorrectes à la fois logiquement et empiriquement. Eliminer des caractères parce qu'ils sont supposés être hautement homoplasiques constitue une entrave *ad hoc* inacceptable au test légitime d'une hypothése. La plupart des traitements statistiques requièrent l'adoption inutile et souvent injustifiée d'hypothèses portant sur les processus évolutifs et la distribution attendue des traits. L'utilisation des seuls caractères et leur apparition sur l'arbre le plus parcimonieux leur sont préférées parce que ce sont les procédures les plus indépendantes d'hypothèses *a priori* et qui demeurent au plus près des données réelles.

INTRODUCTION

The modern systematist has a peculiar place in the natural sciences. His discipline might be characterized as stuffing itself with data while trying to limit general theory. The systematist collects information on morphology, development, genetics, phenology, behavior, ecological associations, biogeographic patterns; indeed, just about anything he can find is potentially useful for deducing hierarchical relationships of species and higher taxa. But as for general theory, there is little more than the process of descent with modification, at least among Hennigian systematists (ELDREDGE & CRACRAFT, 1980: 6). By contrast, nested or intertwined theories about the evolutionary process form the foundation of other fields, such as those concerned with competition or succession ecology, food web structure, sexual selection, or sociobiology. In this light, it is interesting and ironic to see how recent enthusiasm for the primacy of phylogenetic perspectives has revitalized studies of theories about the process of evolution. "Adaptation", one of the central features of Darwinian evolution, has earned the most attention. The papers included in this volume are the result of a symposium that brought together a variety of systematists and others to discuss phylogenetic perspectives on certain evolutionary scenarios. What distinguished this symposium from one that might have been on the systematics of many interesting organisms is that the speakers generally regarded the phylogenetic hypotheses as tools for understanding the process of adaptation (see also EGGLETON & VANE-WRIGHT, 1994; MARTINS, 1996). This method of examination leads to an improvement over the adaptive story-telling of old, but it also carries with it special hazards.

The literature regarding phylogenetic tests of evolutionary scenarios is growing rapidly, but there seems to be a lack of proportional response from a Hennigian perspective on what are the procedures and problems in such an enterprise (but see CARPENTER, 1992b; CODDINGTON, 1994; WENZEL & CARPENTER, 1994). This paper may help to fill that void, or at least draw attention to certain issues that seem to deserve special consideration in this context. One such issue concerns the intent of the study in the first place. A systematist would generally make arguments about various character definitions, states, additivity, etc., and then derive a tree, but in the phylogenetic tests of adaptive hypotheses the process is reversed, producing a tree for the sake of deciding the details of transformations of chosen characters. The phylogeny is only a tool, not a endpoint. The relationships among taxa are the anvil upon which the traits of interest (and any conclusion regarding the process of evolution) are worked into shape. This produces two problems which will be discussed in turn.

HOW GOOD IS THIS TREE?

The first and most obvious problem is that the best phylogenetic hypotheses are generally produced by those who see the phylogeny itself as the end product, but these are often not the people generating the hypotheses for the study in question. Taking care not to sound cynical, we should be cautious when someone who is not very interested in the phylogeny starts producing the hypotheses as a way to get on with his real work (hammering out those interesting traits). Of course, the hypothesis may be quite sound, but one cannot tell simply by reading that the relationships illustrated were derived by analysis with an approved computer program. Even if the primary data are beyond question, issues such as taxon sampling or character coding can have a great impact on tree topology, and even the best computer programs cannot overcome the errors introduced by a researcher's weak understanding of these issues. This problem is particularly important in these days when generating a few DNA sequences for a few taxa and making a tree from them is commonly done by people with no formal education in the sometimes deep problems of phylogenetic reconstruction. The solution, of course, is education.

Support

The simplest rule is to choose the optimal tree for the data in question. This approach has the advantage of logical consistency because all other trees are less well supported. This perspective leads us to the most parsimonious tree, or a consensus of equally most parsimonious trees (FARRIS, 1983; WILEY, 1981; BROOKS & MCLENNAN, 1991)(some non-parsimony based techniques will be discussed in *Resolution*, below). The strict consensus of several equally most parsimonious trees cannot be better resolved than any of the original trees, but the consensus should be considered highly supported anyway because *all* of the clades that appear are included in *all* of the multiple, competing trees. Whenever there is homoplasy, some parts of the tree may be more decisively reinforced by the data than others, but this is not a concern with respect to the optimality criterion of parsimony. By analogy, if we examined the relationship between body length and mass for humans, some points would be better predicted by a straight line than others, but we would nonetheless choose the least squares regression because it is optimal for the complete data set. The homoplasy in a most parsimonious tree can be compared to points that are displaced from a least squares regression line. Each represents an exception to the relationships we expect, but we use our optimum solution anyway.

Even when there is only one most parsimonious tree, measuring support is not straightforward. Ultimately, whether or not a clade is well supported depends on the characters that unite it. While experience and intuition are sometimes sufficient to make assumptions concerning the strength of morphological characters uniting a clade, DNA sequence data usually is not easily evaluated in a straight character by character comparison. What is the relative value of a change from "A" to "G" versus anything else? Also, the relative merits of alternative resolutions of homoplasy in DNA data are often not as logically interpretable as alternative resolutions of morphological homoplasy. The most widely used technique to measure support in this situation is the bootstrap (FELSENSTEIN, 1985), in which the original data are resampled randomly with replacement, new trees are generated from this new matrix, and the clades from the original tree are scored as present or absent on the new tree. This process is repeated and a score is produced for how often such original clades appear, with high scores (100% is perfect) suggesting that the clade is well supported, low scores suggesting it is not. Then clades with low scores (90% might be a cut off point) are dissolved and only strong clades are retained on the "bootstrapped tree".

Objections to the bootstrap include that the statistics generated this way are not comparable to the normal probability estimates people want to use (SWOFFORD *et al.*, 1996: 509); that competition for inclusion in the simulated matrices dictates that uninformative characters (such as autapomorphies) degrade clades when they ought to have no effect (CARPENTER, 1992a; KLUGE & WOLF, 1993); similarly, that strong support in one part of the tree is interpreted as weak support in another part; that asymmetrically branching clades derive higher bootstrap values than symmetrically branching clades even if the relative support for all clades is identical (M. SIDDALL, pers. com.); and that clades with higher empirical content (more taxa) are penalized even when they have the same amount of support as a smaller clade (M. SIDDALL, pers. com.). It is also important to note that the ordinary application of bootstrapping procedures is to derive a confidence interval for the estimate of a parameter of a distribution (MANLY, 1991: 28), and that statistical mean and variance do not apply to unique historical identity of the phylogeny (WENZEL & CARPENTER, 1994: 80 ff.). So, although the bootstrap values mean something, it is not clear what they mean exactly, and it is clear that they do not mean the same thing when compared across different trees. It would seem that these objections would suffice to exile the bootstrap from its current place of honor.

An alternative to the bootstrap is BREMER support (BREMER, 1988), a technique that also gives values for each clade, but these are not intended to serve as statistical confidence tools. Although the calculations are tedious, the principle is simple: start with the most parsimonious tree and then find the shortest trees that do not contain each of the clades in the most parsimonious tree. Because all other trees will be longer, steps will be added when a clade in the most parsimonious tree is broken up. Each clade is then recorded as having support according to the minimum number of steps it costs to break it up. A high value is better support, meaning many characters are *less* parsimonious on the shortest tree that does *not* contain the clade, a low value means few steps are added by interrupting the clade. The problem with documenting BREMER support is that calculations can be prohibitive because of the necessity of calculating so many less-than-optimal trees. The advantage of BREMER support is that values are more easily interpreted than the bootstrap values because they relate directly to how far away from optimal (most parsimonious) the tree would be if it did not contain the clade in question. Interested readers will find a more thorough discussion of the relationship between these measures and short cuts to BREMER support in DAVIS (1995).

A common refrain in studies where trees are used to judge the validity of evolutionary scenarios is "maybe the tree is not good enough", or "maybe some of the branches are not well supported". From a Hennigian perspective, the most parsimonious tree (or consensus of parsimonious trees) is the best tree according to the data. We may want to see values for bootstraps or BREMER support, but we should not change the tree based on these measures.

IS THE TREE GOOD ENOUGH?

The second problem is related to the first one, but it is much older because it surrounds the merits of the phylogenetic hypothesis itself. Phylogenies built for their own sake are usually compared to our previous understanding of the group with the intention of showing that we now know more than we did before. Any increase in understanding is good. However, if such hypotheses are to serve as tools in another context, then a different kind of evaluation is necessary. Gross polytomies may not matter depending on where they fall, so a phylogeny that says very little about relationships may still be "good enough" to answer some questions. Figure 1A illustrates a case in which poor resolution of eight ingroup taxa has no effect on character argumentation. In other cases, character distributions and exact topology can conspire to make even detailed understanding of relationships "not good enough" to answer the question. Figure 1B shows a situation in which we cannot decide whether a character state was derived one, two, three, or four times among seven ingroup taxa, despite having a completely resolved tree.

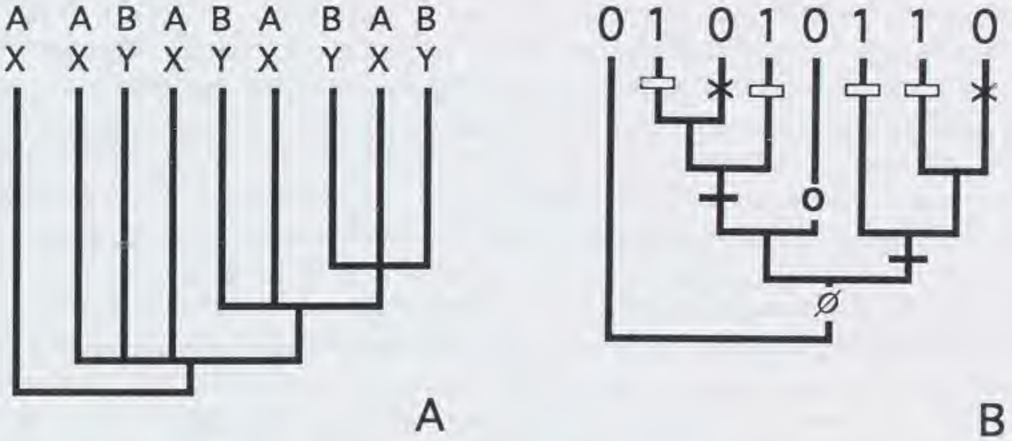

FIG. 1. — **A:** A poor tree that is good enough to answer the question. Two characters exist in two states, (A or B) and (X or Y). There are many ways in which the three polytomies can be resolved and the two characters optimized, but all require at least three separate derivations of the association of AX and BY. This poor tree is good enough to answer the question of multiple origins and convergence. **B:** A complete resolution of seven ingroup taxa that is inadequate to establish character argumentation. The pattern of character states can be explained by four steps, but it is ambiguous as to whether they represent one, two, three, or four derivations of state "1". One scenario has all "1" states derived convergently (open boxes), another scenario has "1" states as two convergent synapomorphies for clades (solid tick marks) followed by two reversals (X). We can combine different halves of these scenarios to get two different ways of having three derivations and one loss (white boxes on one side combined with black ticks and X on the other). Another possibility has "1" as a unique synapomorphy for the ingroup taxa (ø) with three subsequent reversals (X and o). Our interpretation of the significance of character state "1" will change greatly depending on which optimization is chosen, but this excellent tree by itself is not good enough to resolve the problem.

Resolution

One reason strict cladistic techniques seem to be out of favor among many people is that a consensus of equally parsimonious trees often fails to resolve a polytomy. Techniques can improve resolution, such as successive approximation weighting (FARRIS, 1969; CARPENTER, 1988), by which the characters acquire the weights suggested by their congruence with other characters (stability on the cladogram), a new tree is generated with the newly weighted characters, and the process is repeated until a stable solution is obtained. This method provides a way to move from the initial set of trees to the local optimum (which may differ from the global optimum if the original search was too restricted.) Critics have described these methods as circular (SWOFFORD & OLSEN, 1990: 499), but such an appraisal is clearly wrong because the weight of a character is determined by the degree to which it is coherent with *other* characters, and some analyses produce topologies that were not in the original set of trees (BROTHERS & CARPENTER, 1993), hence novel endpoints. The process is recursive, but not circular.

Other reconstruction techniques that do not rely on parsimony, such as neighbor joining or maximum likelihood, may produce one fully resolved tree when parsimony doesn't. Unfortunately, the great shortcoming of these techniques is often the same as their strength: They produce *only* one tree. For example, when data support several alternative relationships equally, neighbor joining techniques (SAITOU & NEI, 1987) choose one arbitrarily based on the order in which taxa appear in the matrix (KLUGE & WOLF, 1993; FARRIS *et al.*, 1996). This means the tree is not strictly determined by the data. Neighbor joining is strictly an algorithm and has no optimality criteria at all, meaning there is no basis for a justification that the tree derived from a given study is somehow the best it can be. In this case, basing additional work on one fully-resolved tree will be at least suspect and perhaps wrong. In contrast, the multiple, equally parsimonious cladograms should not be seen as the failure of a divining rod in the search for water, but rather as the success of a child-proof top on a medicine bottle: if you can't handle this, then you should go no further. Cladistic analyses do not produce a definitive solution when data are ambiguous or lacking, and this result is most consistent with the general principles of the scientific method. Unfortunately, the importance of multiple trees will not be widely appreciated as long as influential publications on phylogenetic methods simply ignore the issue entirely (*e.g.* SWOFFORD & OLSEN, 1990; SWOFFORD *et al.*, 1996).

Judging from several conferences I attended recently, maximum likelihood is rapidly growing to be a popular method of tree-building, and it too can produce resolution or topology that is not present in the data matrix. Many varieties of maximum likelihood estimates are used in modern statistical analyses, and more detailed discussions of their application to phylogenetic hypotheses are available elsewhere (SWOFFORD *et al.*, 1996); here I will approach only the issue of how the tree reflects the data used to produce it. The basic operation, regardless of which maximum likelihood model is used, is that lessons about evolution learned in other studies are applied to the data in question to "improve" our understanding. If other studies have shown that there is an evolutionary bias in the direction of mutation among nucleotide bases, then perhaps we should create a model of the process of evolution that will allow us to build a tree that accounts for this bias. Supporters see this as a strength in that we are using our general knowledge to resolve some local problem, but cladists see this as a very serious flaw. Maximum likelihood models favor some schemes over others *a priori*, and thereby dictate the path that evolution is *expected* to follow (hence, "likelihood"), and then evaluate the degree of likelihood *according to that path*. Although most proponents of maximum likelihood readily acknowledge this, the severity of this shortcoming seems too easily overlooked. Several critics have attacked maximum likelihood methods based on a number of weaknesses (FARRIS, 1986; WENZEL & CARPENTER, 1994), but here I will propose an additional flaw: such methods are self-fulfilling and do not provide independent evidence of their legitimacy. Trees built according to a given model cannot refute the model, and therefore it is clear that the model (which determines to some degree what we will find) is beyond testing. If we decide in advance how evolution is likely to occur, we cannot later declare that we have discovered how evolution occurred. Claims to special knowledge (say, that third positions evolve faster than the rest of the codon) become dictations of pattern, the same pattern that is the point of the inquiry (WENZEL & CARPENTER, 1994). Indeed, arguing in favor of maximum likelihood, SWOFFORD *et al.* (1996: 428, 429) demonstrate that parsimony does not resolve placement of an ambiguous taxon, whereas maximum likelihood does based on the assumption that adding another step to a long branch costs less than adding a step to a short branch (that is, an expectation that more evolution probably occurs on a branch that is already "long" versus one that is already "short"). Thus, the fact that parsimony equivocates when data are equivocal is regarded as a flaw, but that branches that are already short (according to other data) are made to stay short (in the face of new data) is an advantage in maximum likelihood. Such logic would appear to condemn us to a science free from discovery. Some might say that weighting according to branch length is like successive approximations weighting according to homoplasy, but this is correct only superficially. Successive approximations takes over *after* all of the many parsimonious resolutions have been discovered, but maximum likelihood weighs things differently to get the first tree. Whereas parsimony allows all data to compete among themselves in whatever alliances form, maximum likelihood declares *a priori* that some data are favored and others not. Maximum likelihood can never contribute to the business of discovery as meaningfully as does parsimony.

Resisting attacks like that above, some argue that parsimony reconstruction contains its own assumptions. A recent incarnation of this view is that parsimony procedures assume at a minimum that the data are probably generated by a process that would allow parsimony to reconstruct the phylogeny correctly (SWOFFORD *et al.*, 1996: 426). Although this seems almost tautological, it is false nonetheless. Seeking the most parsimonious explanation for data at hand is a fundamental scientific principle and does not represent any statement or assumption about the process that produced the data. There is no assurance that the "truth" is obtained, only that there are infinity *less* parsimonious explanations, so we accept the optimal one and leave questions of "truth" out of it. If the data are misleading, parsimony will yield a misleading answer, and this is as it should be. Garbage in, garbage out. It begs at the margins of clairvoyance to claim that any method can be expected to give good answers from bad data. Strangely, SWOFFORD *et al.* (1996: 426) dismiss in a footnote the argument that parsimony is a fundamental aspect of scientific method, whereas they devote the next 50 pages to non-parsimony methods.

Despite how dismissive they can be about parsimony, proponents of maximum likelihood have been surprisingly forgiving of their method's grave flaws with regard to logical circularity, and quite generous in their acceptance of demonstrations purported to reflect problems in phylogenetic reconstruction. The current school arose, for example, from a model of highly variable rates of evolution with an absolute minimum number of taxa (FELSENSTEIN, 1978). Although some critics have argued that the variable rates of evolution are not an accurate reflection of what most data represent, it is even more obvious that the majority of reconstructions do not consider only four taxa, two of which are very different from the others, and therefore the exercise is highly contrived rather than representative. But, allowing themselves grace, the authors generate data from this model and then show that using this model to interpret the data gives the correct tree whereas parsimony does so less often. Thus, the demonstration is fundamentally circular in addition to being nonrepresentative of general problems in reconstruction. As an example, HULSENBECK (1995) generated artificial DNA sequence data, of lengths from 100 bases to infinity, for four (!) taxa to compare 26 reconstruction techniques. Data were generated from two models, JUKES-CANTOR (using equal base frequencies and one substitution probability) and KIMURA Two Parameters, (using equal base frequencies, but two substitution rates as when transitions and transversion are unequal), and HULSENBECK measured the relative accuracy of different techniques for reconstructing the "true" tree. As an example, when HULSENBECK used the KIMURA model to generate the data, KIMURA gave the best result, thus establishing the perfection of circularity. The closely related JUKES-CANTOR model (see SWOFFORD *et al.*, 1996, p. 434 for relationship among these and six other models) was second best, demonstrating that logic that is nearly circular discovers itself almost as well (see summary in HILLIS *et al.*, 1996, fig. 5). Despite demonstration that the parsimony methods performed well under all circumstances, and despite the fact that parsimony includes no elements of the model used to generate the data, and despite the fact that real evolutionary history is often not stochastic like his models, HULSENBECK concluded that maximum likelihood methods are preferred because they performed best. Other than circular demonstrations of extraordinary problems, there is little to support maximum likelihood as the preferred alternative to parsimony.

Eager to displace parsimony, proponents of maximum likelihood have been slow to produce the necessary studies of how their methods fare when data are *not* derived from the model used to reconstruct them (or more pointedly, not from any model at all), and comparing closely related models (as HULSENBECK did, above) hardly counts as a serious trial. With respect to this charge, and much more relevant to the point of the current paper, maximum likelihood models are based on molecular evolutionary processes that have little relevance to the behavioral, ecological, or life history characters that people want to examine phylogenetically. Ignoring the tautological (maximum likelihood) statement that more evolution is expected to occur on long branches, what assurance is there that the characters associated with the evolution of a new diet, or phenology, or habitat preference should obey the models of evolution based on, say, ribosomal DNA sequences? If defenders argue that maximum likelihood is better at finding the best tree, we should respond that the best tree will be found by including all the data, especially the characters of interest regarding diet, etc. (see *Independence*, below), and then we will be including different kinds of data that can only be combined in a parsimony analysis. Parsimony performs well at reconstructing histories in the absence of any knowledge of the evolutionary model, which is to say it will perform well even when we know no more about our characters than that they have evolved.

Occasionally we will see a presentation in which the researcher has done both parsimony and maximum likelihood analyses to build trees to answer a particular question, perhaps "are polygamous males flamboyant". The two trees usually have similar branching patterns, which increases our feeling that they are valid. The parsimony tree, typically less well resolved, may not be adequate to provide a definitive answer, and so a plea is made to consider the maximum likelihood tree. Such a proposal, although born of necessity, is naive in the extreme. If the parsimony tree shows that there are inadequate data to resolve the problem at this time, then that is the appropriate answer. Defining a tree that is not specified by our data just to use it to answer the external question is irresponsible and dangerous. Certainly, the researcher would not think of making up the other data (declaring, without any information, that a male was both polygamous and flamboyant), so why should he settle for a tree that is not determined by the data?

INDEPENDENCE

There will always be problems regarding the independence of data used in phylogenetic analyses. The most obvious problem is that characters are used to make trees, trees are used to evaluate characters. This problem reduces in part to questions of homology and character definition that are too large for this short paper. Recent treatments from a cladistic perspective are available for the general issue (DE PINNA, 1991), and for morphological (PATTERSON, 1982), behavioral (WENZEL, 1992) or ecological (MILLER & WENZEL, 1995) data. Here we will examine a subset of the general problem: should we include the characters of interest in the phylogeny we will use to evaluate the evolution of those characters? I will argue that we should include the characters, and that intuitive fears of "circularity" are unfounded.

The best test of an evolutionary scenario starts with having the best tree, a tree that relies on as much relevant data as possible. What could be more relevant to the scenario than the characters in question? These characters would be considered useful data if the question was about *other* characters, so why should they be excluded now? This "total evidence" approach (also known as "combined" or "simultaneous" analysis), favored by many strict cladists, is rejected by some because of the idea that it is preferable to have an "independent" test, comparing the traits in question to a tree that was built without their contribution. This widespread opinion has a strong pedigree (CODDINGTON, 1988; BAUM & LARSON, 1991; BROOKS & MCLENNAN, 1992; VANE-WRIGHT *et al.*, 1992) but it is wrong anyway. As DELEPORTE (1993) states nicely, any problem of circularity is restricted to character coding (dictating transformations), not character choice, and therefore flaws are introduced prior to the analysis rather than through analysis itself. This important point deserves more attention than it has received: If our character coding is valid, then combined analysis will introduce no new error. KLUGE & WOLF (1993) argued that the assumption of independence is the same whether comparing across data matrices or within a matrix. No special independence is obtained by partitioning data into different sets, and so there is no circularity created by including all data in a single analysis. Alternatively, if someone can partition any data set into two groups that disagree, are we then obliged to keep them separate forever? Clearly not. Readers uninterested in epistemology might be convinced by reviews of the effects of combined analysis, often producing novel results not found in the partitioned data sets (BARRETT *et al.*, 1991, CHIPPINDALE & WIENS, 1994). Such novel results constitute empirical demonstrations that including all data does not result in circularity.

Even in the spirit of independent sampling, there is no need to exclude the original observations that first suggested a relationship. By analogy, when the first few male students enter a room and sit near the window, and the first few females sit near the door, it is not necessary to exclude these observations from the test to see if this is a general pattern. The students who suggested the relationship are good data, as are the students who have not yet entered the room. With this in mind, consider a more relevant problem: several species of cave crickets are white, blind, and without circadian rhythm. Perhaps they share these traits by descent (the hypothesis of homology), or perhaps these represent independent adaptation (rejection of homology). Combine these data with other characters and allow them to compete among themselves to build a tree. If the cave species come out together, then this means that there is not enough information to reject the proposal that they are alike due to synapomorphy. An obvious question is "what if we get a different phylogeny when the characters of interest are left out?" Then we are still in the position of having an hypothesis of synapomorphy that is *not* rejected, because it still emerges in the combined analysis. The hypothesis of synapomorphy survived the challenges of *all* our data, which is a strong test if we have a lot of other data. In the absence of much other data, we may be subject to statistical Type II error (failure to disprove a false hypothesis), but that is different from circularity.

Some readers may be critical of the cave cricket scenario presented above because synapomorphy is the null hypothesis, and failure to reject the null hypothesis is a weak statement. Such critics may think that adaptation is likely to explain these traits, and that convergence is a more likely explanation than synapomorphy. From this perspective, the null hypothesis would be that the common aspects of the cave crickets are independently derived. Of course, the way to test this proposal is to return all data to the combined analysis and see if we find evidence for synapomorphy. If the commonalties can be plotted as a unique synapomorphy, then we would have to reject the hypothesis of multiple convergent origins. There is no logical way to avoid the combined analysis.

Excluding the characters of interest produces the image of independence when we imagine that the other characters are a repetition of the phenomenon in question. But they are not because they are *other* characters evolved from *other* pressures, and it is rare to see any two characters share precisely the same distribution among all taxa. A tree built on other characters cannot be considered an independent "replicate" of the question we want to examine. Statistically, eliminating the characters of interest is actually more like a jacknife procedure, in which we exclude some data to derive a pattern for the others. Although jacknifing is useful and respected in general, it is not designed for the purpose of providing an estimate of the data that are *excluded*, and it is hard to imagine that anyone would recommend that the way to make sense of certain particular data is to exclude them from the analysis. If the study centers on certain data, then they should be included in the analysis.

Sometimes critics of the cladistic approach want to exclude the characters of interest because these characters are expected to be misinformation, as when independent origins are concealed by extensive convergence. For example, WEST-EBERHARD (1996) endorsed a traditional hypothesis that social parasitism was derived multiply in the genus *Polistes*. Evidence in favor of the hypothesis of separate origins relies on observations of facultative parasitism (stealing nests) in ordinary species. Such variation within original species could become extreme and permanent, eventually resulting in a new species, a social parasite. To defend her position against a strong challenge (CHOUDHARY *et al.*, 1994; CARPENTER *et al.*, 1993), WEST-EBERHARD argued that covariation of morphological traits of interest, (heavy cuticle, square head, powerful mandibles) supply a false indication of synapomorphy (and a single derivation of parasitism) because these traits are expected be convergent adaptations and should have developed independently in each parasitic lineage. Even if parsimonious reconstruction of characters indicates a single origin because the parasites are all closest relatives in a monophyletic clade, this interpretation is ruled out because it conflicts with the theory that parasites should evolve convergently (WEST-EBERHARD, 1996: 315). In this case we must ask ourselves "What information would suffice to disprove the hypothesis of convergence?" If not the patterns among relevant data, then what? There is no good reason to ignore evidence offered by the characters of interest, and exclusion of them constitutes special protection of the hypothesis from a legitimate test. Such a procedure is not within the scientific method.

The example above demonstrates a widespread view opposing combined analysis on the general grounds that "bad" data will lead us away from the better answer we could have had with the "good" data alone. Most of these authors are proponents of maximum likelihood, which may explain why I do not agree with them. For example, HULSENBECK *et al.* (1996) offer a brief review and promote a method (HULSENBECK & BULL, 1996) to identify "pathological" data during reconstruction of the phylogeny for the four (!) taxa lizard, alligator, chicken, and mouse. They argue that 18S rRNA sequence is more in conflict with other genes than is expected by pure sampling error, and that we should then keep the 18S data separate from the other data and come up with explanations about why the 18S data are different. Aside from questions about how relevant such a study is to real phylogenetic reconstruction, one might well ask what we expect to achieve by combining (good) partitions that agree and separating (bad) partitions that don't (see also BROWER, 1996, for additional criticism of the method). The struggle between "good" and "bad" data is hard for me to understand because of the difficulty in knowing which data are "good" and which are "bad". As argued above, the business of discovery seems to be associated with patterns that were not expected (hence "discovery") which means they are "bad" data according to the maximum likelihood method. More to the point, when carefully-examined "adaptive" traits are plotted on a phylogeny, it seems that they are relatively well behaved (see other papers in this volume), which demonstrates that adaptation does not confuse the larger pattern of evolution, and that cautious homology statements recognize independent origins of similar traits. DELEPORTE (1993) is supported empirically: if we are careful judging our characters, we do not create problems in reconstruction. In situations where evidence indicates separate origins for a character of interest, it is appropriate to reexamine that character and see what differences might be found in the similar (but non-homologous) states. Recursive examination is the best way to make use of information that was not available at the beginning of the study. From the point of view of the phylogenetic analysis itself, such reappraisal commonly identifies homoplasy that is due to coding problems rather than real evolutionary problems, which is a necessary step to better understanding (MILLER & WENZEL, 1995). From the point of view of character of interest, reappraisal deepens our understanding of the nature and limits of similarity among separately derived states. Character reexamination should be an integral part of all phylogenetic studies, whether focused on ecological transitions or not, or the process of discovery will be crippled pointlessly.

STATISTICS

Many people become scientists because they like the idea of discovering things, knowing the answers to important questions, and deciding which of competing ideas is true. Nature, however, does not always let us find a clear path to truth. Some solutions to this problem were developed by early geneticists, who were forever presented with measurement errors of various sorts, and biology in general followed them to become more statistical. Statistics is not only a way to summarize data, but a way to decide in favor of A or B when faced with some uncertainty. It is only logical that the science of statistics has become associated with the uncertainty of phylogenetic reconstruction as it has with nearly all else in biology. The problem with this new development is that a phylogeny is unique, it is an historical event that happened once and cannot be resampled. Whereas statistical analysis is good for predicting what sort of distribution or expectation we should get from tossing a coin (because we can sample and resample those events), it is poor for saying whether the fourth toss was heads or tails (because that happened only once). History is not a sampling problem, and efforts to make it become one are pseudoscientific at best. In the context of phylogenetic tests of adaptation, WENZEL & CARPENTER (1994) and LEROI *et al.* (1994) discussed the inapplicability of certain procedures to phylogenetics, and readers would do well to review those arguments. What is most relevant here is that a researcher can always make a phylogenetic test or reconstruction highly statistical, but would that actually make it better? The cladistic viewpoint is to rely upon the primary observations as much as possible.

"Statistics As Truth" is a motto of many researchers who despise the idea of indecision. There may be an overt advertisement for an artificial decision rather than an examination of natural data, as with this endorsement of arbitrary values substituted for real observations: "... it does allow analysis of the data now, rather than waiting for actual phylogenetic information to become available" (GARLAND *et al.*, 1992: 19; see WENZEL & CARPENTER, 1994, for treatment of other examples). Other researchers substitute statistics to produce an image of quantification when none is necessary. LOSOS (1992) presented a statistical approach in which lizard ecomorphs from different island radiations were separated by principle component analysis (PCA) on morphometric values. Then a phylogeny was used to reconstruct the PC scores for hypothetical ancestors and decide what ecological transitions occurred in the separate clades. For this to be valid, the covariance matrix of all characters must remain the same through evolutionary time, which appears to me to rule out a lot of evolution, such as the evolution of species (whether ancestral or terminal) that are allometrically unique with respect to other species. This assumption should be a source of concern, but for now let us overlook it. In two radiations (Jamaica versus Puerto Rico) the transitions through PC space were from a twig-resting form to a tree-crown form to a trunk form, with grass and bush forms derived most apically, and it is reported that this is significant at $P<0.04$ (LOSOS, 1992: 412). It is not discussed whether there would have been a different answer if we just reconstructed "twig", "crown", "trunk", and "grass", which seems to be the first thing to try. Such a reconstruction would be based on the actual data of interest and would be free of assumptions about the process we are trying to discover. If the patterns the author found are not supported by optimization of "twig", "crown", "trunk", and "grass", but rather due to the statistical reconstruction of hypothetical ancestors, then there is no evidence supporting his theory other than that it is consistent with the assumptions he made about the evolutionary process; that is to say it is not supported by the primary data. If the patterns are supported by the primary data, then there is no point in all the statistical manipulation. An example of completely needless statistics giving a clearly wrong answer can be found in COGNATO *et al.* (1997). Examining the sex pheromone mixture of ten species of beetles, the authors asked if the pheromone variation is congruent with phylogenetic history or not. The test consisted of a resampling of data from the original matrix (excluding the characters of interest) and evaluating the average homoplasy for these when plotted on a cladogram. When the average homoplasy for the characters of interest proved to be higher than the average for the other characters, it was decided that they were *not* congruent with phylogenetic history. Yet, four of nine pheromone components plot with no homoplasy, and three require only one additional step (COGNATO *et al.*, 1997: fig. 1). Only two components seem to be "poor", trans-verbenol and verbenone, both of which are suspected to be artifacts of unrelated chemical processes. So, whereas the answer from plotting data on a tree is that at least four and perhaps seven components are congruent with phylogeny, the authors confused themselves into rejecting that proposal. Fortunately, some researchers are content to plot data on a tree and let these patterns speak for themselves with no appeal to probability values (BASOLO, 1996; EMERSON, 1996; MCLENNAN, 1996).

EPILOGUE

It is inspiring to see that biologists of all sorts are interested in using phylogenetic perspectives to illuminate their evolutionary studies. We must see that rigorous scientific methods are satisfied during the difficult task of discovering phylogenetic patterns. Cladistic perspectives have been ignored sometimes because general eagerness to include phylogenies has produced a demand for single, well resolved trees, and unambiguous tests that are easily interpretable. Unfortunately, these things cannot be delivered simply because we desire them. Many methods or opinions that enjoy broad support among widespread researchers were not derived from careful consideration of their consequences or implications about how we do science. Hennigian methods remain the most well-founded and will produce the results best suited to serve as the foundation of future research.

ACKNOWLEDGEMENTS

I thank P. GRANDCOLAS and the Muséum National d'Histoire Naturelle for hosting the symposium from which this paper came. I thank J. M. CARPENTER, P. GRANDCOLAS, M. E. SIDDALL, D. W. WENZEL, and an anonymous referee for comments on the manuscript. S. A TEALE kindly shared unpublished work.

REFERENCES

BARRETT, M., DONOGHUE, M. J. & SOBER, E., 1991. — Against consensus. *Systematic Zoology*, **40**: 486-493.

BASOLO, A. L., 1996. — The phylogenetic distribution of a female preference. *Systematic Biology*, **45**: 290-307

BAUM, D. A. & LARSON, A., 1991. — Adaptation reviewed: a phylogenetic methodology for studying character macroevolution. *Systematic Zoology*, **40**: 1-18.

BREMER, K., 1988. — The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, **42**: 795-803.

BROWER, A., 1996. — Combining data in phylogenetic analysis. *Trends in Ecology and Evolution*, **8**: 334-335.

BROOKS, D. R. & MCLENNAN, D. A., 1991. — *Phylogeny, Ecology, and Behavior*. Chicago, University of Chicago Press: 1-434.

BROTHERS D. J. & CARPENTER, J. M., 1993. — Phylogeny of Aculeata: Chrysidoidea and Vespoidea. *Journal of Hymenopteran Research*, **2**: 227-302

CARPENTER, J. M., 1988. — Choosing among multiple equally parsimonious cladograms. *Cladistics*, **4**: 291-296.

CARPENTER, J. M., 1992a. — Random Cladistics. *Cladistics*, **8**: 147-153.

CARPENTER, J. M., 1992b. — Comparing methods. [Review of P.H. Harvey and M. D. Pagel, 1991, The Comparative Method in Evolutionary Biology, Oxford University Press.] *Cladistics*, **8**: 191-195.

CARPENTER, J. M., STRASSMANN, J. E., TURILLAZZI, S., HUGHES, C. R., SOLIS, C. R. & CERVO, R., 1993. — Phylogenetic reationships among paper wasp social parasites and their hosts (Hymenoptera: Vespidae; Polistinae). *Cladistics*, **9**: 129-146.

CHOUDHARY, M., STRASSMANN, J. E., QUELLER, D. C. , TURILLAZZI, S. & CERVO, R. 1994. — Social parasites in polistine wasps are monophyletic: implications for sympatric speciation. Proceedings of the Royal Society of London, B, **257**::31-35.

CHIPPINDALE, P. T. & WIENS, J. J., 1994. — Weighting, partitioning, and combining characters in phylogenetic analysis. *Systematic Biology*, **43**: 278-287.

Cognato, A. I., Seybold, S. J., Wood, D. L. & Teale, S. A., 1997. — A cladistic analysis of pheromone evolution in *Ips* bark beetles (Coleoptera: Scolytidae). *Evolution*, **51**: 313-318.

Coddington, J. A., 1988. — Cladistic tests of adaptational hypotheses. *Cladistics*, **4**: 3-22.

Coddington, J. A., 1994. — The roles of homology and convergence in studies of adaptation. *In*: P. Eggleton & R. Vane-Wright, *Phylogenetics and Ecology. Linnean Society Symposium Series. n°17*. London, Academic Press: 53-78

Davis, J. I., 1995. — A phylogenetic structure for the monocotyledons, as inferred from chloroplast DNA restriction site variation, and a comparison of measures of clade support. *Systematic Botany*, **20**: 503-527.

Deleporte, P., 1993. — Characters, attributes and tests of evolutionary scenarios. *Cladistics*, **9**: 427-432.

De Pinna, M. C. C., 1991. — Concepts and tests of homology in the cladistic paradigm. *Cladistics*, **7**: 367-394.

Eggleton, P. & Vane-Wright, R., (eds), 1994. *Phylogenetics and Ecology. Linnean Society Symposium Series. n°17*. London, Academic Press: 1-376.

Eldredge, N. & Cracraft, J., 1980. — *Phylogenetic Patterns and the Evolutionary Process*. New York, Columbia University Press: 1-349.

Emerson, S. B., 1996. — Phylogenies and physiological processes - the evolution of sexual dimorphism in southeast asian frogs. *Systematic Biology*, **45**: 278-290.

Farris, J. S., 1969. — A successive approximations approach to character weighting. *Systematic Zoology*, **18**: 274-385.

Farris, J. S., 1983. — The logical basis of the phylogenetic system. *In*: N. I. Platnick & V. A. Funk, *Advances in Cladistics, Vol. 2: Proceedings of the Second Meeting of the Willi Hennig Society*. New York, Columbia University Press: 7-36.

Farris, J. S., 1986. — On the boundaries of phylogenetic systematics. *Cladistics*, **2**: 14-27.

Farris, J. S., Albert, V. A., Källersjö, M., Lipscomb, D. & Kluge, A. G., 1996. — Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, **12**: 99-124.

Felsenstein, J., 1978. — Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, **27**: 401-410.

Felsenstein, J., 1985. — Confidence limits on phylogenies: an approach utilizing the bootstrap. *Evolution*, **39**: 783-791.

Garland, T., Harvey, P. H. & Ives, A. R., 1992. — Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Sytematic Biology*, **41**: 18-32.

Hillis, D. M., Mable, B. K. & Moritz, C., 1996. — Applications of molecular systematics: the state of the field and a look to the future. *In*: D. M. Hillis, C. Moritz & B. K. Mable, *Molecular Systematics*. Sunderland, Massachusetts, Sinauer Associates: 515-543

Hulsenbeck, J. P., 1995. — Performance of phylogenetic methods in simulation. *Systematic Biology*, **44**: 17-48

Hulsenbeck, J. P. & Bull, J. J., 1996. — A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology*, **45**: 92-98.

Hulsenbeck, J. P., Bull, J. J. & Cunningham, C. W., 1996. — Combining data in phylogenetic analysis. *Trends in Ecology and Evolution*, **11**: 152-158.

Kluge, A. G. & Wolf, A. J., 1993. — Cladistics: What's in a word? *Cladistics*, **9**: 183-199.

Leroi, A. M., Rose, M. R. & Lauder, G. V., 1994. — What does the comparative method reveal about adaptation? *The American Naturalist*, **143**: 381-402.

Losos, J., 1992. — The evolution of convergent structure in Caribbean *Anolis* communities. *Systematic Biology*, **41**: 403-420.

Manly, B. F. J., 1991. — *Randomization and Monte Carlo Methods in Biology*. New York, Chapman & Hall.

McLennan, D. A., 1996. — Integrating phylogenetic and experimental analyses: The evolution of male and female nuptial coloration in the stickleback fishes (Gasterosteidae). *Systematic Biology*, **45**: 261-277.

Martins, E. P., (ed), 1996. — *Phylogenies and the Comparative Method in Animal Behavior*. New York, Oxford University Press: 1-415.

Miller, J. S. & Wenzel, J. W., 1995. — Ecological characters and phylogeny. *Annual Review of Entomology*, **40**: 389-415.

Patterson, C., 1982. — Morphological characters and homology. *In*: K. A. Joysey & A. E. Friday, *Problems in Phylogenetic Reconstruction*. London, Academic Press: 21-74.

Saitou, N. & Nei, M., 1987. — The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**: 406-425.

Swofford, D. L. & Olsen, G. J., 1990. — Phylogeny reconstruction. *In*: D. M. Hillis & C. Moritz, *Molecular Systematics*. Sunderland, Massachusetts, Sinauer Associates: 411-501.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M., 1996. — Phylogenetic inference. *In*: D. M., Hillis, C. Moritz & B. K. Mable, *Molecular Systematics*. Sunderland, Massachusetts, Sinauer Associates,: 407-514.

VANE-WRIGHT, R. I., SCHULTZ, S. & BOPPRE, M., 1992. — The cladistics of *Amauris* butterflies: congruence, consensus and total evidence. *Cladistics*, **8**: 125-138.

WENZEL, J. W., 1992. — Behavioral homology and phylogeny. *Annual Review of Ecology and Systematics*, **22**: 361-381.

WENZEL, J. W. & CARPENTER., J. M., 1994. — Comparing methods: adaptive traits and tests of adaptation. *In*: P. EGGLETON & R. VANE-WRIGHT, *Phylogenetics and Ecology. Linnean Society Symposium Series. n°17*. London, Academic Press: 79-101.

WEST-EBERHARD, M. J., 1996. — Wasp societies as microcosms for the study of development and evolution. *In*: S. TURILLAZZI & M. J. WEST-EBERHARD, *Natural History and Evolution of Paper Wasps*. London, Oxford University Press: 290-317

WILEY, E. O., 1981. — *Phylogenetics: The Theory and Practice of Phylogenetics Systematics*. New York, John Wiley & Sons: 1-439.